

dilibri – ein einheitliches Digitalisierungsportal für Rheinland-Pfalz

Barbara Koelges, Hagen Reinstein, Elmar Schackmann und Hans-Ulrich Seifert

Ausgangslage

Digitale Publikationen gewinnen mehr und mehr an Bedeutung. Für viele ist die Informations- und Literatursuche im Netz an der Tagesordnung, und auch Wissenschaftler sind in ihrer Arbeit zunehmend auf elektronische Publikationen angewiesen. Diese bieten neben der schnellen Verfügbarkeit komfortable Recherchemöglichkeiten.

Angesichts dieser Entwicklung wächst der Druck auf Bibliotheken ihre umfangreichen Altbestände an Büchern, Zeitschriften und historischen Dokumenten in digitaler Form zu Verfügung zu stellen.

In einigen großen Bibliotheken, so z. B. in Göttingen und München, wurden in den letzten Jahren mit finanzieller Unterstützung der Deutschen Forschungsgemeinschaft (DFG) Digitalisierungszentren aufgebaut. Für viele kleinere Bibliotheken ist jedoch der Kosten- und Personalaufwand für den Aufbau einer eigenen Digitalisierungsabteilung zu hoch.

Das Landesbibliothekszenrum Rheinland-Pfalz (LBZ) versteht sich als Servicezentrum für alle Fragen der Informations- und Literaturvermittlung und sieht es als seine Aufgabe an, rheinland-pfälzische Bibliotheken bei dieser in Zukunft immer wichtiger werdenden Tätigkeit zu unterstützen. So wurde die Idee einer landesweiten, gemeinsam nutzbaren Plattform für Digitalisierung geboren. Die Universitätsbibliothek (UB) Trier, die im Bereich Digitalisierung bereits einige Erfahrungen gesammelt hatte, konnte als Kooperationspartner für das Projekt dilibri gewonnen werden.

Dilibri (www.dilibri.de) (Abb. 1) ist die digitalisierte Sammlung von landeskundlichen Werken zu Rheinland-Pfalz sowie von Beständen aus rheinland-pfälzischen Bibliotheken. Dilibri bietet rheinland-pfälzischen Bibliotheken, die ihre Bestände digitalisie-

ren, eine zeitgemäße, gemeinsame Plattform.

Grundlagen und Regeln

In erster Linie ist dilibri ein Portal für urheberrechtsfreie landeskundliche Literatur über Rheinland-Pfalz und in Rheinland-Pfalz erschienene Werke. Das Portal ist aber auch offen für andere Bestände rheinland-pfälzischer Bibliotheken, soweit diese eine Digitalisierung für sinnvoll erachten. Die beteiligten Bibliotheken entscheiden selbst, welche Bestände sie vorrangig in dilibri einbringen. Nach der Pilotphase von September 2007 bis Dezember 2007, in der zu Testzwecken Werke mit verschiedenen Publikationsformen und unterschiedlichen Schrifttypen eingestellt wurden (historische Zeitschriften, Adressbücher, landeskundliche Monographien, illustrierte Werke und Schulschriften), digitalisiert das LBZ in Koblenz seit Beginn des Echtbetriebs u.a. die Koblenzer Adressbücher in Zusammenarbeit mit dem Stadtarchiv Koblenz, Werke aus der Sammlung illustrierter Rheinbücher des 19. Jahrhunderts, daneben aber auch landeskundliche Zeitschriften und Monographien, die für die historische Forschung von Bedeutung sind. Die UB Trier bringt in erster Linie ältere historische Zeitschriften und Monographien aus dem Trierer Raum ein, während die Stadtbibliothek Mainz als weiterer Partner damit begonnen hat, Mainzer Fastnachts-

zeitungen des 19. Jahrhunderts zu digitalisieren.

Um Doppeldigitalisierung zu vermeiden, werden die vorgesehenen Werke im Zentralen Verzeichnis digitalisierter Drucke (ZVDD), in Google books und bei Drucken aus dem 19. Jahrhundert im Katalog der Bayerischen Staatsbibliothek geprüft.

Als Grundlage für die Scannerstellung dienen die „Praxisregeln im Förderprogramm Kulturelle Überlieferung“¹ der Deutschen Forschungsgemeinschaft (DFG). Alle Inhalte von dilibri stehen unter einer Creative Commons Lizenz.²

Software

Nach mehrjähriger Marktsichtung und nach Abwägung des Für und Wider der auf dem Markt befindlichen Produkte entschloss sich das LBZ in Absprache mit der UB Trier zur Anschaffung des Software-Produkts „Visual Library“ der Firma semantics Kommunikationsmanagement GmbH aus Aachen.

„Visual Library“ ist ein Modul des Produkts Scantoweb, welches von der Walter Nagel GmbH & Co. KG angeboten wird. „Visual Library“ besteht aus den beiden Komponenten „Visual Library Server“ und „Visual Library Manager“. Die Software deckt den gesamten Bearbeitungsworkflow ab: den

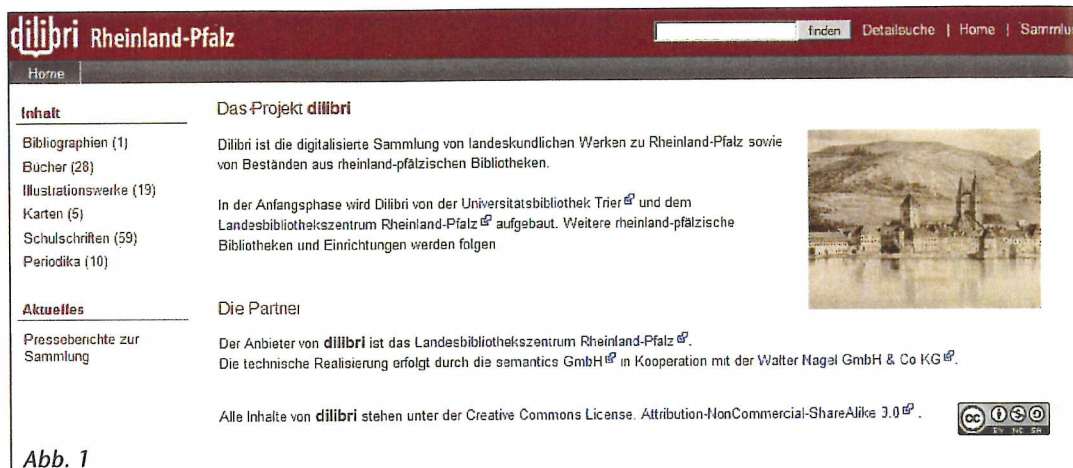
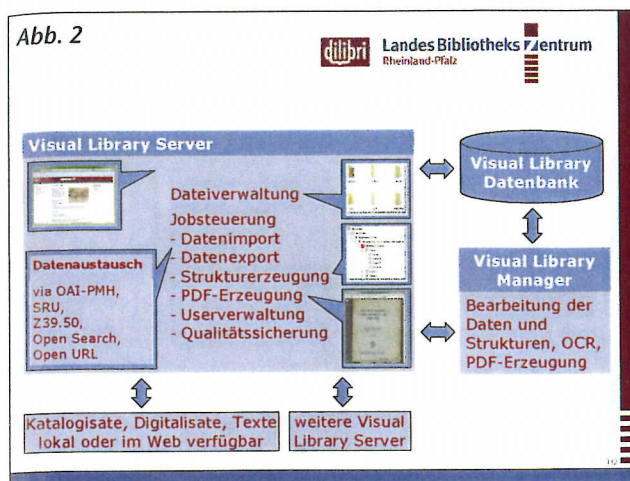


Abb. 1

1 http://www.dfg.de/forschungsfoerderung/formulare/download/12_151.pdf (Stand 13.8.2008)

2 <http://de.creativecommons.org/> (Stand 13.8.2008)

Abb. 2



Import der Scans, die Überwachung und Steuerung der Verarbeitungsprozesse, die automatisierte Verknüpfung der Katalogdaten mit den digitalisierten Dokumenten, die Erschließung und Weiterverarbeitung der Digitalisate sowie die Bereitstellung des Inhalts in einem eigenständigen Portal oder integriert in die eigene Webseite.

Die Bearbeitung der Digitalisate erfolgt mittels der Client-Anwendung „Visual Library Manager“. Zu deren optional kombinierbaren Modulen gehört unter anderem ein umfassendes Qualitäts- und Workflow-Management, welches den Anwendern einen detaillierten Blick auf den aktuellen Projektstatus bietet. Integriert ist auch ein OCR-Modul, mit dessen Unterstützung Digitalisate über Schriftenerkennung automatisch ausgelesen, die Inhalte gespeichert und als Volltext der späteren Recherche zur Verfügung gestellt werden. Ein XML-Editor zur Korrektur von erkannten Texten sowie ein PDF-Generator runden das Angebot ab.

Der „Visual Library Server“ kümmert sich um die Verwaltung und Verarbeitung der anfallenden Dateien, den Datenim- und -export, den Datenaustausch und die Präsentation der dilibri-Homepage, während der „Visual Library Manager“ einen gut bedienbaren Zugriff zur Bearbeitung dieser Daten bietet und die Strukturierung der Daten sowie die OCR- und PDF-Erstellung auf dem Server anstößt (Abb. 2).

Warum hat sich das LBZ nicht für eine auf dem Markt befindliche „Open-Source-Lösung“ bzw. „Freie Software“³ entschieden? Generell gibt es im LBZ und auch in der UB Trier Sympathie für den Einsatz von Open-Source-Lösungen. Als Beispiel sei Typo3 genannt, das als Content-Management-System die Grundlage für die Erstellung der Homepage der UB Trier und des LBZ ist.

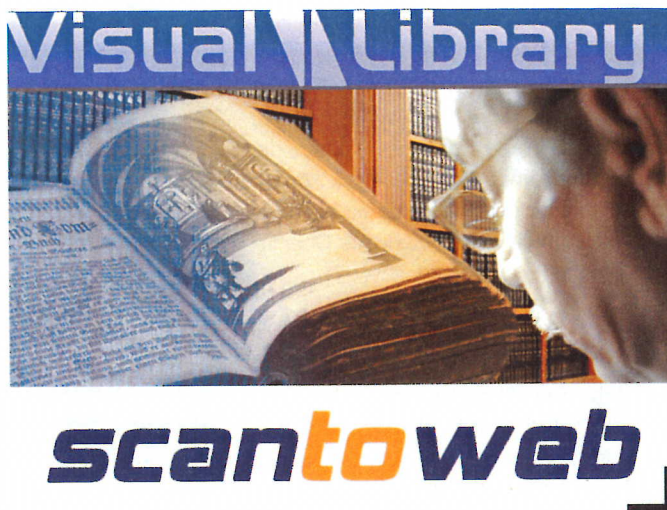
Wer schon einmal eine komplexe Homepagestruktur mit Typo3 erstellt hat, weiß, dass dies ohne „man-power“ und „Know-how“ nicht so einfach zu leisten ist. Besonders Ersteres ist in der IT-Abteilung des LBZ nur in sehr begrenztem Maße vorhanden bzw. in laufenden Arbeiten und Projekten gebunden.

Bereits 2004 wurde in einer Diplomarbeit⁴ evaluiert, ob sich die Open-Source Digital Library Software „Greenstone“⁵ für den Einsatz in der Rheinischen Landesbibliothek (RLB), jetzt ein Standort des LBZ, eignet. Nach dieser Evaluation hat sich die RLB gegen den Einsatz ausgesprochen.

³ Definitionsabgrenzung „Open Source“ und „Freie Software“ s. dazu: http://de.wikipedia.org/wiki/Open_Source bzw. http://de.wikipedia.org/wiki/Freie_Software (Stand 12.8.2008)

⁴ Tomaschoff, Andre: Die Einsatzmöglichkeiten der Software Greenstone zur Web-Präsentation von Digitalisierungsprojekten in der Rheinischen Landesbibliothek. Darmstadt, Fachhochsch., Fachbereich Informations- und Wissensmanagement, Diplomarb., 2005

⁵ Greenstone Digital Library Software <http://www.greenstone.org/> (Stand 12.8.2008)



Ihre digitale Bibliothek
in sicheren Händen

Wir begleiten Sie

vom
Scannen

bis ins
WEB-Portal

Besuchen Sie uns:
Frankfurter Buchmesse
15.-19. Oktober 2008
Halle 4.2 N445

Walter Nagel
innovative Organisationslösungen

Walter Nagel GmbH & Co KG
Tennenweg 7
33729 Bielefeld

Tel. 0521 92479 -0

Die Entscheidung für das kommerzielle Produkt „Visual Library“ wurde also ganz bewusst getroffen, da wir hier ein Produkt vorfanden, das bereits viele Automatisierungsmöglichkeiten integriert hatte, die wichtigsten Schnittstellen bedienen konnte und sowohl für den Bearbeiter als auch für den Nutzer intuitiv zu bedienende Oberflächen bot.

Technische Infrastruktur

Am Koblenzer Standort des LBZ wurde ein Linux-Server (mit Betriebssystem Open Suse) installiert und von der semantics GmbH mit ihrem Softwarepaket „Visual Library Server“ bestückt.

Daneben wurde für das Digitalisierungsportal ein dediziertes Speichersystem aufgestellt. Dabei handelt es sich um ein flexibel erweiterbares Storage Area Network von Qlogic, das in der Grundausstattung mit einer Speicherkapazität von 5,7 Terabyte ausgestattet wurde. Ein vorhandener Terminalserver wird genutzt, um den Bearbeiterzugriff auf die Client-Software „Visual Library Manager“ sowohl innerhalb des LBZ als auch per Remote-Zugriff den Partnern UB Trier und Stadtbibliothek Mainz zu ermöglichen. Dieser Terminalserverzugriff ist sinnvoll, da bei der Bearbeitung der Scans diese auf dem Server verbleiben können und dadurch die zu bearbeitenden Digitalisate nicht zwischen den beteiligten Institutionen durch die Datenleitung gezwängt werden müssen, was bei der Größe der Digitalisate vor allem eine Zeitverzögerung während der Bearbeitung bedeuten würde.

Geschäftsgang

Die Aufnahme von Dokumenten in dilibri erfolgt anhand eines klar definierten Workflows. Die Scans werden per File Transfer Protocol (FTP)-Verbindung in das Speichersystem geladen. Durch die Übergabe der Identifikations-Nummer der Verbunddatenbank des Hochschulbibliotheksverbunds Nordrhein-Westfalen (hbz) werden beim Import die bibliographischen Metadaten der gedruckten Werke über die Z39.50-Schnittstelle aus der hbz-Verbunddatenbank übernommen und im „Visual Library Server“ mit den Digitalisaten verknüpft. Durch diese automatisierte Nutzung bereits vorhandener Katalogdaten entfällt eine eigene Katalogisierung in dilibri weitestgehend. Bei mehrbändigen Werken und bei Zeitschriftenbänden können bibliographische Zusatzinformation händisch eingebracht werden. Die UB Trier und das LBZ sind Katalogisierungsteilnehmer des hbz-Verbundes. Inzwischen nutzt die Stadtbibliothek Mainz als Teilnehmerin des Hessischen Bibliotheks-

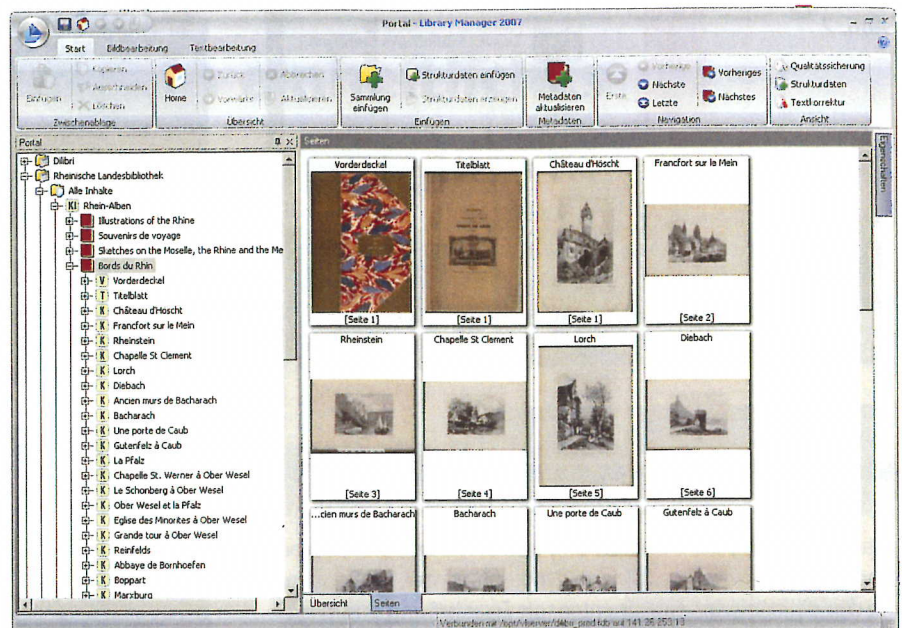


Abb. 3

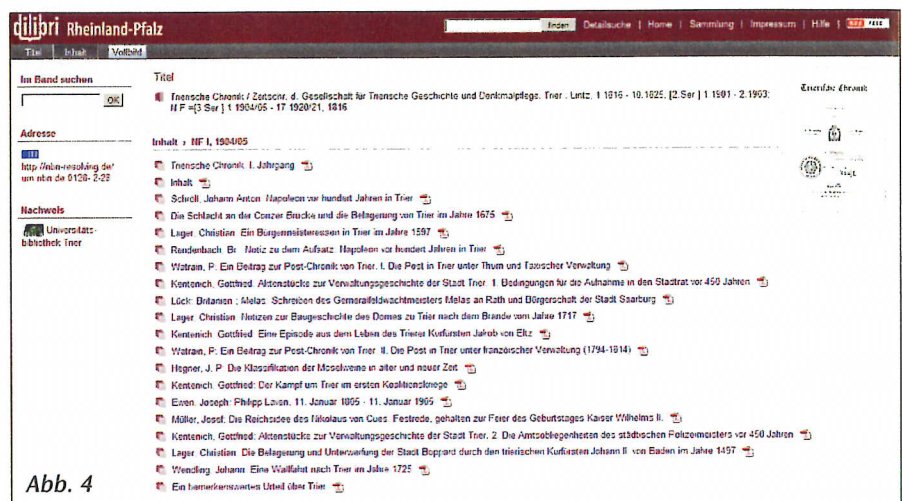


Abb. 4

Informationssystem (HeBIS) eine entsprechende Schnittstelle, die die Metadaten aus dem HeBIS-Verbund den Mainzer Digitalisaten zuordnet.

Nun können die geladenen Objekte im „Visual Library Manager“ (VLM) weiterverarbeitet werden. Es besteht die Möglichkeit, über das Modul Qualitätssicherung zu prüfen, ob die Scanqualität den selbst aufgelegten Standards genügt. So wird u.a. überprüft, ob die Scans die gewünschte Auflösung und Größe haben. Die Qualitätssicherung bietet ebenfalls einen detaillierten Blick auf den aktuellen Status der Bearbeitung (z.B. in Nachbearbeitung, Abgenommen, ...).

Jetzt können die Objekte strukturiert werden, die einzelnen Scans also den Strukturdaten vom Buchdeckel über das Titelblatt und das Inhaltsverzeichnis bis hin zu den Kapiteln zugeordnet werden (Abb. 3). Dadurch kann das Inhaltsverzeichnis digital nachgebildet und dem Benutzer eine einfach zu handhabende und schnelle Navigation im Objekt bereit gestellt werden. Die

dem System zu Grunde liegenden Strukturelemente stammen aus dem DFG-Viewer-Projekt⁶.

Anschließend werden durch einen Knopfdruck die OCR-Erkennung und die PDF-Erzeugung gestartet. Es besteht die Möglichkeit, jedem Objekt, jedem Strukturelement oder falls nötig jeder Seite die Schriftart und Sprache zuzuordnen. So gibt es bereits jetzt Objekte, die sowohl Teile in Frakturschrift besitzen als auch Kapitel in lateinischer, französischer und spanischer Antiqua. Sind die entsprechenden Erkennungsmodi den einzelnen Teilen zugeordnet, kann für das Gesamtobjekt die OCR-Erkennung gestartet werden. Im Hintergrund läuft dazu die ABBYY FineReader Engine. Spannend war die Frage, wie gut die Ergebnisse der Frakturerkennung sein würden: Je nach Vorlage und verwendetem Schrifttyp ist die OCR-Frakturerkennung sehr unterschiedlich, aber doch so gut, dass sich die dilibri-Partner

⁶ <http://dfg-viewer.de/profil-der-strukturdaten/> (Stand 12.8.2008)

sowohl bei der Recherche in den einzelnen Dokumenten als auch im gesamten dilibri-Pool einen erkennbaren Mehrwert versprechen.

Es ist möglich, die OCR-Ergebnisse nachzubearbeiten, was im augenblicklichen Stadium illusorisch ist. Es bleibt anderen Projekten jedoch unbenommen, eine händische Bearbeitung von einzelnen Objekten (etwa bei den Adressbüchern) vorzunehmen.

Ist das Objekt fertig bearbeitet, so kann es für die Einsicht im Web-Portal www.dilibri.de freigeschaltet werden. Mit der Freischaltung ist die dilibri-Neuerscheinung ebenfalls sofort als RSS-Feed abrufbar. Eine beim Scan-Import bereits generierte URN (Uniform Resource Name) steht nun für den Harvester des URN-Resolvers der Deutschen Nationalbibliothek bereit. Über die OAI-Schnittstelle⁷ werden die neuen Objekte z.B. über OAIster⁸ eingesammelt und können dort recherchiert werden.

Die dilibri-Daten können in verschiedenen Metadatenformaten bereitgestellt werden: im Metadata Encoding & Transmission Standard (METS)⁹, gemäß dem Metadata Object Description Schema - MODS¹⁰ oder im MARCXML-Format¹¹. In absehbarer Zeit werden die Katalogisate der dilibri-Objekte als Sekundäraufnahmen in den hbz- bzw. HeBIS-Katalog eingespielt und damit in den lokalen Katalog übertragen werden. Zeitschriften und Zeitungen werden weiterhin in der Zeitschriftendatenbank bzw. in der Elektronischen Zeitschriftenbibliothek (EZB) aufgenommen und gelangen über den Weg des Verbundkataloges in den lokalen OPAC.

Webpräsentation

Die Webpräsentation von dilibri bietet eine intuitive Navigation innerhalb eines Objektes, eine Anzeige der Digitalisate in verschiedenen Größen und meist einen Textestieg über Strukturdaten (also eine Art Inhaltsverzeichnis) für das jeweilige Objekt (Abb. 4). PDF-Dateien des gesamten Objektes oder von einzelnen Kapiteln stehen in der Regel zum Herunterladen bereit. Eine Volltext- und Metadatensuche über den gesamten Bestand wird ebenso geboten wie die Suche innerhalb eines Objektes. Im Hintergrund läuft Lucene, die Open-

Source-Java-Bibliothek zum Erzeugen und Durchsuchen von Text-Indizes¹². Es gibt weiterhin sowohl die Möglichkeit des Browsers im Gesamtbestand als auch das Angebot, die neu eingestellten Digitalisate per RSS-Feed abzufragen.

Für die Nutzung von dilibri sehr erfreulich ist das gute Ranking der Objekte in Google und Co: So finden sich viele Werke auf Platz 1 der Google und Yahoo-Suchergebnisse (Bsp: Rheinische Geschichtsblätter)¹³.

Wie geht's weiter?

Dilibri wurde inzwischen allen interessierten rheinland-pfälzischen Bibliotheken vorgestellt und die Teilnahmebedingungen erläutert. Die Stadtbibliothek Mainz arbeitet bereits mit und der Standort Pfälzische Landesbibliothek des Landesbibliothekszen-trums wird die nächste Teilnehmerbibliothek sein. Der größere Teilnehmerkreis wird dazu beitragen, die Attraktivität von dilibri für private wie wissenschaftliche Nutzer der Digitalisierungsplattform weiter zu steigern.

Nun kommt es darauf an, dilibri weiter zu optimieren: Die in dilibri befindlichen Metadaten müssen in die jeweiligen Kataloge und Verzeichnisse wie z. B. ZVDD transportiert werden – die notwendigen Schnittstellen sind vorhanden. Weiterhin soll der Automatisierungsgrad des Workflows vergrößert werden, insbesondere bei der Kopplung der Scannerstellung und der Vergabe der Strukturdaten liegt noch Optimierungspotential. Zurzeit finden hier erste vielversprechende Tests statt. So kann der Scanoperator bereits Strukturelemente wie Titelblatt, Inhaltsverzeichnis, Kapitel etc. per Tastendruck definieren. Beim Import der Scandaten wird die Strukturierung in „Visual Library“ umgesetzt, sodass der Strukturierungsvorgang vereinfacht werden kann.

Eine weitere Zukunftsaufgabe ist das Thema Langzeitarchivierung. Dieses ist weder finanziell noch technisch abschließend geklärt.

Auch die Frage der automatisierten Vernetzung der in dilibri gewonnenen Digitalisate mit bibliographischen Nachweisinstrumenten wie der Rheinland-Pfälzischen Bibliographie (<http://www.rpb-rlp.de/>) stellt ein dringendes, aber nicht einfach einzulösendes Desiderat dar. Und umgekehrt ist zu überlegen, ob einschlägige landeskundliche Literatur zu Rheinland-Pfalz, die in anderen

Projekten wie Google Booksearch oder dem (leider ausgelaufenen) Microsoft Internet Archive digitalisiert wurde, in dilibri nachgewiesen werden sollte.

Sukzessive soll sich dilibri zu einem gemeinsam genutzten, landesweiten Digitalisierungsportal für landeskundliche Werke zu Rheinland-Pfalz bestückt aus unterschiedlichen Einrichtungen des Bundeslandes entwickeln. Ohne finanzielle Unterstützung aus zentralen Landesmitteln wird ein solcher Service sich freilich nicht dauerhaft etablieren und weiterentwickeln lassen.

Mit dilibri steht Rheinland-Pfalz bereits jetzt eine einheitliche Plattform für Bibliotheken zur Verfügung, um landesweite Projekte im Bereich der Digitalisierung zu realisieren. Dies ist ein großer Vorteil gegenüber anderen Bundesländern, in denen größere Bibliotheken zwar ebenfalls ihre Bestände digitalisieren, aber mit unterschiedlichen Software-Lösungen und Programmen arbeiten, so dass deren Nutzer in verschiedenen Systemen recherchieren müssen.

■ DIE AUTOREN

DR. BARBARA KOELGES

Referentin für Öffentlichkeitsarbeit des LBZ, Referentin für Handschriften und alte Drucke am LBZ-Standort Rheinische Landesbibliothek
koelges@lbz-rlp.de



DR. HAGEN REINSTEIN

Fachreferent an der UB Trier und dilibri-Betreuer
reinstein@uni-trier.de



ELMAR SCHACKMANN

Systembibliothekar am LBZ Rheinland-Pfalz in Koblenz und dilibri-Koordinator
schackmann@lbz-rlp.de



DR. HANS-ULRICH SEIFERT

Fachreferent an der UB Trier, Leiter der Abt. Digitale Medien
seifert@uni-trier.de



Landesbibliothekszen-trum / Rheinische Landesbibliothek
Bahnhofplatz 14
56068 Koblenz

⁷ <http://www.openarchives.org/> (Stand 18.8.08)

⁸ <http://www.oaister.org/> (Stand 13.8.08)

⁹ <http://www.loc.gov/standards/mets/> (Stand 13.8.08)

¹⁰ <http://www.loc.gov/standards/mods/> (Stand 13.8.08)

¹¹ <http://www.loc.gov/standards/marcxml/> (Stand 13.8.08)

¹² <http://lucene.apache.org/> (Stand 13.8.08)

¹³ <http://www.google.de/search?q=rheinische+geschichtsbl%C3%A4tter> (Stand 13.8.2008) bzw. <http://de.search.yahoo.com/search?p=rheinische+Geschichtsbl%C3%A4tter> (Stand 13.8.2008)